2016

Misunderstanding Reliability

Klaus Krippendorff
University of Pennsylvania, kkrippendorff@asc.upenn.edu

Misunderstanding Reliability

Misunderstanding Reliability

Klaus Krippendorff
The Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, USA

On *ResearchGate*, I received an anonymous invitation to comment on Feng (2015).

I was surprised that *Methodology* published Feng's paper, hesitated to write a comment as it might be perceived as self-serving, but decided to accept the editors of this journal's invitation. I share their belief in dialogue as a way of advancing scientific methodology and hope readers will consider my comments seriously and critically.

I contend Feng discusses reliability measures with seriously mistaken conceptions of what reliability is to assure us of, starting with the very first sentence of his paper (p. 13):

"Intercoder reliability assesses the degree of agreement or consensus in the ratings given by judges . . ."

I consider this definition unsatisfactory. In English, "reliability" stands for the ability to rely on the use of something. In content analysis, to which Feng's paper refers, reliability is the ability to rely on data that are generated to analyze phenomena a researcher intends to study, theorize, or use in pursuit of practical decisions. As such, reliability is independent of whether data result from mechanical measuring devices or human observers or coders. In the social sciences, the phenomena of analytical interest tend to be encoded in written documents, socially or culturally happenings, or interviews of political actors, rarely in physical form that mechanical devices could capture. To turn social phenomena into analyzable data, employment of literate and socioculturally astute coders is indispensable. The explicit coding instructions that coders are asked to apply are the link to know what the resulting data are about. Data have to link the phenomena one wishes to study to conclusions drawn from them.

One demonstration of the reliability of data involves comparing them to a known standard. Substantial agreement with that standard establishes their accuracy.

In the absence of a standard, the more common way to assess the reliability of data is by demonstrating that independent replications of applying written coding instructions by qualified human coders come to the same conclusions. Substantial agreement among such replications allows us to infer the extent to which data can be considered as reliable surrogates for phenomena of analytical interest, phenomena that coders saw, judged, categorized, valued, and recorded as analyzable data. Demonstrating replicability can provide researchers with the needed assurance that the data they are analyzing are not or are only minimally polluted by irrelevant variation or noise.

So, Feng's starting definition of reliability puts the cart before the horse. Reliability is not a measure of the agreement or consensus among coders; it needs to be conceptualized the other way around. We need to measure the extent of agreement among independent replications in

order to estimate whether we can trust the generated data in subsequent analyses. Replicability certainly is affected by the degrees to which given coding instructions are flawlessly communicable to and applicable by the coders in the researchers' employment as well as by equally qualified individuals that other researchers might hire when using these instructions in subsequent research efforts. Measures of replicability need to assure researchers that the variance of the generated data is explainable by the differences that coders detected among the phenomena they recorded.

Feng's narrow focus on measuring agreement and consensus among coders fails to provide the assurances that researchers need in order to proceed with analyzing the data in place of the phenomena they wish to explore. He never mentions that data have to be about something of analytical interest and the methodological consequences of analyzing them. This goes along with Feng's consistent preoccupation with the cognition of coders. He insists on reliability measures that reflect the difficulties that coders experience in reaching consensus among them. Regarding this misconception, Feng follows the footsteps of Zhao, Liu, and Deng (2012) who introduced psychological terminology into the discussion of reliability. This led them to such outrageous claims that the so-called chance-corrected agreement coefficients falsely assume coders to be dishonest. Not only is the ability to rely on data independent of whether the data result from mechanical measuring devices, are generated by human coders, or found, the claim that current agreement coefficients have built dishonesty into their mathematical forms is simply unsustainable. These authors have joined hands by simplistically equating reliability with coders' ease of categorizing units of analysis, and unreliability with coders' difficulties to do the same.

Yes, there are phenomena that are easy to code and others that are difficult to code. And yes, it is often so that the difficulties that coders experience when coding result in unreliable data. Indeed, when coders are uncertain on which category or value accurately describes a particular unit of analysis and disagree, phenomena-unrelated variance or noise enters and pollutes the data, effectively reducing the information that reliable data could provide to an analysis of the phenomena under scrutiny.

Experiences of coding difficulties challenge the designers of coding instructions to do better. But measures of these difficulties and how much work has to be done to improve the written instructions have little to do with the ability to trust the generated data in subsequent analyses. If Feng wanted to define measures of coding difficulties, he should not have talked about reliability and get entangled in numerous confusions that permeate his paper.

For example, Feng claims that the percent agreement measure is good for variables that are easy to code while Krippendorff's α should be applied only to variables that are difficult to code. It is totally unclear what Feng means by "good for." But it is clear that he recommends selecting reliability measures according to an undefined sense of the difficulties that coders are having. Offering researchers the choice of agreement coefficients that reflect their sense of coder difficulties opens the door for researchers to present the quality of their data in deceptively favorable lights.

I guess Feng, Zhou et al. allow themselves to be misled by taking the adjective "inter-coder" or "inter-rater" of agreement or consensus literally, privileging the cognitive ability of human

coders instead of the reproducibility of coding instructions across a population of coders. Feng questions my saying that coders should be treated interchangeably. Interchangeability of coders is justified when accepting a notion of reliability that has to do with the quality of generated data, not with what or who generates them. While the need to delegate the process of generating data to human coders is undisputed whenever the phenomena of interest are linguistically or socially complex, their cognitive abilities, the strategies they employ when coding, and how they interpret the categories defined in the coding instructions are a side issue to whether the resulting data can inform a subsequent analysis about the phenomena these coders saw, interpreted, described, and recorded as data.

Feng cites Klemens (2012) who

"argued that the error model of (Cohen's, 1960) κ, (Scott's, 1955; Fleiss, 1971) π, and (Krippendorff's, 2013) α is implausible as they assume that coders work by flipping coins all the time. . . . echoed by Feng (2013b), who empirically tested the error models of the three indices." (p. 14)

I have yet to see a coder who assigns categories by flipping a coin. In my experiences, coders always try to do the best they can even in the face of unworkable coding instructions.

Feng presents the common form of so-called chance-corrected agreement coefficients for nominal data by:

$$\left\{\begin{matrix}\pi\\ \kappa\\ S\\ \alpha\\ \mathrm{AC1}\end{matrix}\right\} = \frac{Po - Pe}{1 - Pe} = 1 - \frac{1 - Po}{1 - Pe} = 1 - \frac{Do}{De}\,. \quad (1)$$

He acknowledges that the observed agreements Po (and I would include their maximum Po = 1 in the denominator), common to all of these coefficients, are corrected by different kinds of chance or expected agreements Pe. But his conception of coder difficulties leads Feng to discuss them only by projecting "different assumptions about the coding distributions" on them. In fact, neither these coefficients nor their authors make such assumptions. Pe defines the zero points of the scales that these coefficients provide. In particular:

π – corrects the proportion of observed agreement Po by the probability Pe of agreement observable in the absence of any statistical relationship between the data and the phenomena they are claimed to be about. This probability is obtained by pairing all collectively used categories in the data without replacement. The distribution of the collectively used categories is the best estimate of the characteristics of the population of data. By pairing them without replacement, π's Pe is computed for infinite sample sizes.
κ – does not correct Po as π does. Instead, κ's Pe is the probability of agreement observable when two coders are statistically independent of each other. Statistical independence omits disagreements among the two coders' predilections for particular categories (evident in unlike marginal frequencies of contingency matrices). Feng claims that κ assumes such disagreements. In fact, κ is immune to them.

S – (Bennett et al., 1954) corrects Po by the probability Pe = 1/ k, where k is the number of categories a coding instruction makes available for coding. Unlike what Feng claims, S does not assume anything about the frequencies with which coders categorize the phenomena in question, S simply has no place for them.
α – corrects Po just as π does but for finite not infinite sample sizes. Its Pe is the proportion of agreements between all possible pairs of categories or values used without replacement. It follows that α's Po and Pe are proportions, not a mixture of proportions and probabilities.
AC_1 (Gwet, 2002, 2008)'s correction will be discussed below.

Evidently, what the literature calls "chance-agreement" has rather different meanings in these coefficients. Feng is not the only author who is utterly disinterested if not unaware of what these coefficients respond to or ignore. He complains that chance corrections do not reflect the difficulty level of coding tasks, and claims that the Pe of these coefficients influence the measured agreements "abnormally" without defining what normal could possibly be. I consider it a serious omission not to examine their differences carefully before judging them.

We know that unreliable data (i.e., data whose representation of the phenomena of interest is polluted by noise) reduce the resulting correlations and associations, show findings to be less statistically significant than they would be had data been reliable, or misguide an analysis to invalid findings. Reliability tests need to warn researchers of such possibilities. Gustafson (2004) explored the effects of measuring errors and misclassifications on a variety of statistics used in epidemiology. He did not propose a reliability coefficient but demonstrated how unreliable data affect the outcomes of analyzing them. To show these effects, Gustafson had to distinguish the information in data on which an analysis can rely for computing potentially valid conclusions from the error variance which is unrelated to the phenomena of interest. This is precisely what a reliability coefficient has to do without being sidetracked by irrelevant individual coder-related issues.

In this regard, the Pe in Cohen's κ can lead to particularly misleading assessments of data reliability. As already mentioned, κ's Pe is the agreement when coders are statistically independent of each other. It fails to account for inter-coder disagreements on the marginal frequencies of contingency tables, adding them to the agreements instead! This has the well-recognized effect of κ punishing coders for agreeing on their marginal frequencies and rewarding coders for deviating in their uses of categories. κ's Pe resembles the computation of associations, contingencies, and correlation coefficients which have no reason to regard coders as interchangeable. Consequently, when such disagreements occur, κ inflates agreement beyond valid estimates of data reliability. Surprisingly, Feng notes that κ's Pe is unlike that for π or α's but does not seem to care what it omits and when it distorts indications of data reliability. It is simply incorrect to say as Feng does that κ assumes coders to disagree on their marginal uses of categories. κ simply fails to recognize them as disagreements.

Feng mentions my considering AC_1 as "odd" and dismissing it as a reliability coefficient. Contradicting my assessments, Feng claims that:

"The chance agreement of AC_1 is positively correlated with the difficulty level of coding tasks, hence is actually a better agreement index than Krippendorff's α, π, and κ." (p. 14)

That AC_1 corrects the observed percent agreement Po by a measure of chance or expected agreement is simply false. Feng may have noticed that it is something else by observing that "the calculation of chance agreement of Gwet's AC_1 is exactly opposite to that of Scott's π," but he fails to recognize the implications of flipping these quantities. In fact, AC_1 corrects Po not by anything resembling expected or chance agreement. For dichotomous data, it corrects Po by the expected disagreement, De = 1 - Pe, which is the quantitative complement of the expected agreement Pe. Accordingly:

$$AC_1 = \frac{Po - (1-Pe)}{1-(1-Pe)} = \frac{Po - De}{Pe}! \qquad (2)$$

I am not the only one for whom the proportion of *observed agreement* minus *expected disagreement* to the expected agreement does not make sense. I should mention that for numbers of categories larger than two, AC_1 correction of Po is more complex but far from easier interpretable. Feng claims but gives no evidence of how AC_1 correction correlates with the difficulty of coding. He echoes Gwet's claims of AC_1's superior qualities, but whatever it measures, it cannot possibly be called a chance-corrected agreement coefficient. For dichotomous data, AC_1 is a chance corrected agreement coefficient if and only when De = Pe. This rather unique condition limits AC_1 to a situation of two replications with two categories occurring with equal frequencies. In that case $AC_1 = \pi = \kappa = S$. Feng notes favorably that AC_1's values exceed those of the other chance-corrected agreement coefficients, but does not or cannot explain what AC_1's inflations are due to, what they mean. Because of its generally higher values, AC_1 is attractive only to those who, like Feng, have no clue of why AC_1 behaves that way and how its values relate to reliability. Actually, "odd" is a very generous attribute. Feng's promotion of AC_1 encourages an uninterpretable measure.

Feng also refers to Klemens (2012) who proposed a coefficient called P_I of the ratio of information in agreements to the total variance in the data – measured in terms of Shannon's entropies. I have bad news for Feng and Klemens. Well before I developed α, I had proposed the very same ratio (Krippendorff, 1971) but subsequently found that the unequal granularities of entropy measures in these ratios did not allow comparing entropies across different sample sizes, unequal number of coders, etc. I realized that Shannon's entropies are measures of distributions and influenced by the number of categories with above-zero frequencies. They just can't do the job. This is why I shifted my attention to measuring reliabilities in pairs, with variance analysis-like measures, but continued to express reliability as the proportions of information in the data about the phenomena of analytical interest. Variance measures do not have the problems I experienced with entropies.

As might be seen from the above explorations, evaluating agreement coefficients by a detailed analysis of what their mathematical structures are capable of responding to is more convincing, at least to me, than simulating their values for a variety of numerical examples (Feng, 2013b) or quoting unsubstantiated opinions and popular consensus by anonymous sources. Feng writes:

"Although it has been a consensus that percent agreement generally overestimates reliability in that it does not make allowance for chance agreement, but it is not considered as a misuse if used

for nominal scaled codings. The rationale is that Zhao et al. (2012) and Feng (2013b) found that percent agreement is not a bad index when the coding task is easy." (p. 17)

The references to Zhao et al. and Feng featured intuitions of what is good and bad but without offering quantitative criteria. Mathematically, percent agreement Po:

- Can indicate the reliability of data if and only when two conditions are met: (i) agreement is without exceptions and (ii) data exhibit sufficient variance for Po to have statistically significant consequences. In the absence of variance, data cannot convey information about differences among phenomena of declared analytical interest, contingencies, correlations, and similar statistics cannot be computed, and by mathematical necessity, agreement is always 100%. Without variance any agreement coefficient should be silent regarding the reliability of data.
- On the other end of the scale, Po's lowest value, 0%, is statistically unlikely. It can occur when coders either agree to disagree on categorizing each unit, which violates the requirement of replications to be independent of each other; or they unknowingly apply different coding instructions to the same phenomena, which renders the resulting data uninterpretable. It follows that Po = 0 is not interpretable in terms of the replicability of data.
- The range of Po's values over and above the absence of any meaningful relationship between data and the phenomena of interest is a function of the number of categories available for coding. For two categories, percent agreement may measure anywhere between 50% and almost 100%, for three between 33% and also almost 100%, etc.

These mathematical facts do not change whether one likes Po or not, whether Po is popular or not, or whether coding is difficult or easy. Feng's observation that Po "generally overestimates reliability" is correct. But highlighting this property does not discourage its use. As shown above, percent agreement has nothing definite to say about the trustworthiness of data except when Po = 100% and data exhibit sufficient variance.

Feng's Table 2 states several equivalences among reliability coefficients. Unfortunately, these equivalences leave much to be desired:

It is simply false that for two coders and nominal data Krippendorff's α = Fleiss' π = Scott's π. Unlike both πs, α does not compute its expected agreement as if data were of infinite size. Only when data become very large do π and α become equal. Regarding multiple coders, Fleiss' π requires complete replications by a fixed number of coders, α accepts variable numbers of contributors to the reliability data, can cope with missing values, and adjusts to the observed sample sizes, Fleiss' π does neither. To the best of my knowledge, the weighted κ (Cohen, 1968) and Lin's (1989) ρ_r are not defined for rank ordered data, as Feng claims, at least not by customary conceptions of rank order differences; for example, in Spearman's rank order correlation coefficient ρ (rho). Feng did not mention α's applicability to rank ordered data. For interval data, α is not quite equal to the intraclass correlation coefficient ICC. As already noted in conjunction with nominal data, the interval α too responds to actual sample sizes while ICC assumes them to be infinite. Finally, Feng does not distinguish between interval and ratio data (Stevens, 1946). Applying coefficients that are designed to respond to interval differences to data

in the form of proportions and ratios omits the reliabilities that are particular to ratio data. Incidentally, CCC[1] has been shown to be equal to my (Krippendorff, 1970) α for interval data.

Not only because of Feng's introduction of coder difficulties in the decision tree of Figure 1, the above inaccuracies and omissions render Feng's decision tree of questionable value.

By invoking Feinstein and Cicchetti's (1990) so-called paradox of "high (percent) agreement but low kappa," Feng joins Feinstein, Cicchetti, and Zhao et al. who fail to recognize the already mentioned importance of variance in data to speak confidently about their reliability. Instead, they assume that the more intuitive percent agreement Po should positively correlate with any measure of the reliability of data. In his paper, Feng confuses my uncontroversial proposition that statistically rare phenomena require larger sample sizes to support statistically significant inferences about data reliability with efforts (not mine) to avoid this so-called paradox when he claims that he:

"(Feng, 2013a) has empirically demonstrated that Krippendorff (2011)'s excuse as well as ad-hoc procedure of testing information adequacy is unfounded." (p. 19)

I checked on what Feng's quote could possibly be referring to and found even more misunderstandings. In the paper that Feng cites, I had proposed a way to estimate the sample sizes needed for observed agreements to serve as indicators of data reliability at a chosen level of statistical significance. Based on such estimates, I also suggested a coefficient able to indicate the extent to which α-agreement could be trusted to measure data reliability. Feng mistakenly assumed that adequate sample sizes would eliminate Feinstein and Cicchetti's paradox. However, if α is low or zero, increasing the sample size may not change its value. Adequate sample sizes have to do with the statistical certainty of α's interpretation. Feng's comment reveals his inability to distinguish between the two issues.

In opposition to the alarm raised by Feng, Gwet, and Zhao et al. holding up Feinstein and Cicchetti's paradox, I contend that low α-values properly indicate that the generated data are insufficiently informative about the phenomena to be analyzed. The phrase "high (percent) agreement low kappa" (α included) seems paradoxical only if one privileges, as these authors do, percent agreement and believes that reliability must correlate with any valid agreement coefficient, even when data do not provide information about the phenomena to be analyzed. Without variance, analyses of data cannot contribute to statistically significant findings comparisons, contingencies, correlations, and other descriptive statistics. Under these conditions it is perfectly justified that low variances drive chance-corrected agreement coefficients toward zero. That these authors do not seem to see this relationship attests to a conception of reliability without variance which divorces this conception from the larger context of research in which reliability plays a decisive role.

Feng's content analysis of the use of reliability in the communication literature is not the first. Relative to previous findings, the frequencies he reports in his Tables 4 and 5 are not entirely surprising. However, the fact that percent agreement P_o was still the most popular measure, closely followed by π, demonstrates the need to make the community of communication researchers aware of what available agreement coefficients actually indicate. Unfortunately,

Feng's favorable mention of P_o and silence about the defective κ is counterproductive in this regard. Although Feng describes several misuses that everyone can easily agree with, his misconceptions overshadow his conclusions.

Feng tested the reliability of his assistant's categorizations by replicating 30% of his or her data and computing seven agreement coefficients. The results are a mixed bag but worth a brief comment. The chance-corrected coefficients α, π, and κ are all smaller than the coefficients that I gave sufficient reasons to distrust: S, AC_1, P_o (and I_r, not discussed here). By presenting these measures as alternatives, Feng gives the impression that researchers can pick what suits them without informing them about what these choices imply. I hope my foregoing comments will enable researchers to be more selective than Feng seemingly is. Feng's lack of attention to details is also evident in his computations of agreement coefficients. We do not have access to his reliability data but one has to be wary of the accuracy of the calculated agreements when the reported numerical results of the four study types he reports in his Table 3 violate the quantitative relationship between α, π, and κ. Stated above albeit in words:

$\alpha \geq \pi \leq \kappa$, (3)

$\pi = \kappa$ if and only when coders perfectly agree on the marginal distribution of categories. $\pi < \kappa$ when coders disagree on the marginal distribution of categories showing κ to inflate agreement by mistakenly adding this disagreement to κ. The reported differences demonstrate κ's small but noticeable bias: $\pi = 0.58 < \kappa = 0.59$.

$\alpha = \pi$ if and only when samples are large, ideally infinite in size (Disagreements on the marginal distributions are registered by both alike). When samples are small $\alpha > \pi$. But Feng's results show $\alpha = 0.44$ to be smaller not larger than $\pi = 0.58$. This is mathematically impossible.

Let me make five recommendations beyond those previously published (Hayes & Krippendorff, 2007; Krippendorff, 2004, 2013) in the hope that they discourage unsound reliability testing of the kind encountered above:

- Publish or make available everything needed to replicate the data making process elsewhere. This ought to include the coding instructions, training manuals, and criteria for selecting coders from a population of qualified candidates. Also assure critical readers that replications were made independent of each other and surreptitiously derived consensus could not have influenced the observed agreement.

- Make sure that the size of the reliability data is adequate to lead to statistically significant inferences about the reliability of subsequently analyzed data. Whether it is enough to replicate only a fraction of the data whose reliability is at stake and/or by the minimum of two coders needs to be tested, not left open. Apply only agreement coefficients to the replications that indicate (i) the extent to which differences among the phenomena of analytical interests are represented in the variance of data, and (ii) data contain sufficient amounts of relevant information for their subsequent analysis – agreement coefficients not to exceed the level of measurement underlying the generated data:

nominal < ordinal < interval < ratio

- Scott's π for two coders, nominal data (categories), and large sample sizes.
- Fleiss' π for a fixed number of coders, no missing categories, and large sample sizes.
- Krippendorff's α for any number of coders, whether categories or values are missing, any one level of measurements, and all sample sizes.

- Report the reliabilities of all variables of the data examined. If the reliability of a whole system of variables is at issue, report the smallest reliability among the variables actually utilized. (Average reliabilities should not be computed.)
- Report the confidence intervals of the agreement coefficients and their Type I errors above a chosen minimally acceptable level of agreement. That minimum should be a function of the costs of drawing statistically insignificant or wrong conclusions from (even only minimally) unreliable data.

To sum up, while I do not think there is a last word on reliability, there is always room for improvements, I was compelled to comment on Feng's paper, including the works of others he drew into it, because they promote what I consider serious misconceptions in the literature on reliability. I fear that not understanding these misconceptions can encourage unsound empirical scholarship. What is needed are explorations of how measurable unreliabilities in data affect the results of diverse analytical techniques, similar to Gustafson's (2004) investigations, even if it means defining reliability coefficients that are specialized for particular analytical efforts.

Note:

1. This was intended to refer to Lin's (1986) ρ_r.

References:

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, 18, 303–308.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Cohen, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213–220.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43, 543–549.

Feng, G. C. (2013a). Underlying determinants driving agreement among coders. *Quality & Quantity*, 47, 2983–2997.

Feng, G. C. (2013b). Factors affecting intercoder reliability: A Monte Carlo experiment. *Quality & Quantity*, 47, 2959–2982.

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology – European Journal of Research Methods for the Behavioral and Social Sciences*, 11, 13–22. doi: 10.1027/1614-2241/a000086

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378–382.

Gustafson, P. (2004). *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian Adjustments*. New York, NY: Chapman & Hall/CRC Press.

Gwet, K. L. (2002). Kappa statistics is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-Rater Reliability*, 1, 1–5.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61, 29–48.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1, 77–89.

Klemens, B. (2012). Mutual information as a measure of intercoder agreement. *Journal of Official Statistics*, 28, 395–412.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30, 61–70. doi: 10.1177/300001316447003000105

Krippendorff, K. (1971). Reliability of recording instructions: Multivariate agreement for nominal data. *Behavioral Science*, 16, 222–235.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30, 411–433. doi: 10.1111/j.1468-2958.3042004.tb00738.x

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Measures and Methods*, 5, 93–112. doi: 10.1080/19312458.2011.568376

Krippendorff, K. (2013). *Content analysis. An introduction to its Methodology* (3rd ed.). Thousand Oaks, CA: Sage. Replacement of Section 12.4 in its 2nd printing. Retrieved from http://www. asc.upenn.edu/usr/krippendorff/U-alpha.pdf

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*, 255–268.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321–325.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.

Zhao, X., Liu, J. S., & Deng, K. (2012). Assumptions behind intercoder reliability indices. In C. T. Salmon (Ed.), *Communication Yearbook* 36 (pp. 419–480). New York, NY: Routledge.

Klaus Krippendorff
The Annenberg School for Communication
University of Pennsylvania
Philadelphia, PA
USA
klaus.krippendorff@asc.upenn.edu